# Modelling interdependencies between fish species in the North Atlantic

## A Bayesian machine learning approach to predictive biomass and recruitment models

Larsson, Aron Mikael
UNIVERSITY OF HELSINKI, FACULTY OF BIOLOGICAL AND ENVIRONMENTAL SCIENCES

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

| | | |
|---|---|---|
| Tiedekunta – Fakultet – Faculty<br>Faculty of biological and environmental sciences | Laitos – Institution– Department<br>Environmental sciences | |
| Tekijä – Författare – Author<br>Aron Mikael Larsson | | |
| Työn nimi – Arbetets titel – Title<br>Modelling interdependencies between fish species in the North Atlantic<br>A Bayesian machine learning approach to predictive biomass and recruitment models | | |
| Oppiaine – Läroämne – Subject<br>Environmental change and global sustainability | | |
| Työn laji – Arbetets art – Level<br>Masters thesis | Aika – Datum – Month and year<br>04/2021 | Sivumäärä – Sidoantal – Number of pages<br>41 |

Tiivistelmä – Referat – Abstract
The science of fish stock assessment is one that is very resource and labor intensive, with stock assessment models historically being based on data that causes a model to overestimate the strength of a population, sometimes with drastic consequences. The need of cost-effective assessment models and approaches increases, which is why I looked into using Bayesian modeling and networks as an approach not often used in fisheries science. I wanted to determine if it could be used to predict both recruitment and spawning stock biomass of four fish species in the north Atlantic, cod, haddock, pollock and capelin, based on no other evidence other than the recruitment or biomass data of the other species and if these results could be used to lower the uncertanties of fish stock models. I used data available on the RAM legacy database to produce four different models with the statistical software R, based on four different Bayes algorithms found in the R-package *bnlearn*, two based on continuous data and two based on discrete data.
What I found was that there is much potential in the Bayesian approach to stock prediction and forecasting, as our prediction error percentage ranged between 1 and 40 percent. The best predictions were made when the species used as evidence had a high correlation coefficient with the target species, which was the case with cod and haddock biomass, which had a unusually high correlation of 0.96.
As such, this approach could be used to make preliminary models of interactions between a high amount of species in a specific area, where there is data abundantly available and these models could be used to lower the uncertanties of the stock assessments. However, more research into the applicability for this approach to other species and areas needs to be conducted.

Avainsanat – Nyckelord – Keywords
Bayes, Bayesian networks, machine learning, R, bnlearn, fisheries, recruitment, biomass, prediction, forecasting

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Sakari Kuikka

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

# INTRODUCTION

Fisheries research is a very observation-based, resource and labor-intensive form of science. As the need for information of fisheries around the world increases, the annual cost for agencies and organizations in these fields increases. The International Council for the Exploration of the Sea (ICES) has an annual budget of around 6,5 million euros (ICES annual budget 2019), while the European Commission Executive Agency for Small and Medium-sized Enterprises (EASME) reports that the annual cost for research surveys at sea under the Data Collection Framework Multi Annual Plans costs around 33 million euros annually (EASME 2018). This underlines the need for effective use of existing data to decrease the uncertainties of stock assessment.

Stock strength and recruitment have historically been subject to overestimation (Walters & Maguire 1996). These assessments have sometimes been based solely on catch-effort data without biomass, fishing mortality and reference point assessments. These make stock forecasts unreliable, as they do not consider changes in environmental factors, fishing mortality and selectivity (Froese et al 2017). By implementing a new way of assessing fisheries data with Bayesian networks, we may lower the uncertainties tied to traditional assessment methods. Uncertainties and estimation errors may get very costly if they cause failures in fisheries management and stock overexploitation occurs. This may cause a moratorium to be declared on the fisheries. When this happened, for example, with the northern cod stock in Newfoundland in Canada in 1992, some 30.000 fishermen and plant workers were put out of work, with a benefits program TAGS costing the government 1.9 billion dollars (Heritage 2008).
With a total catch of cod in 2017 being 868.000 metric tons just in the area researched in this thesis, northern Atlantic ICES subareas 1 and 2, the economic impacts of the global fisheries are significant.

Fisheries research has two different centers of attention when assessing stocks. These are: 1) impacts of fisheries and 2) impacts of environmental factors on fish

populations. Fisheries management needs causality in research, as it is what guides the management of fisheries (impact of action on stocks and fisheries) and legislation connected to it (Sugihara et al 2012). However, in recruitment research and in stock assessment models, causality is not necessarily needed, as the underlying cause for the change in the stock is not the focus of management in e.g. total allowable catch, as the only relevant knowledge there is the quantity of individuals in a stock.
There may be underlying effects, such as environmental factors, that cause the changes and correlations in and between species, but knowing the causes of these variables (so called hidden variables which are not in observational data) are not necessarily needed, as the correlation between the species' biological responses may be mapped out and modeled without deeper understanding of the mechanisms. This way we may forecast population and recruitment changes and lower the uncertainty of predictions without the need of complex models and pricey surveys of both stocks and environmental factors.
Correlational analysis is often used to map out where causality might be present. Correlation alone does not give indication that the two events are connected to each other in a causal way. However, we want to test if making predictions based on correlational data could be a useful tool for fisheries management, as it is cost-effective and may reduce uncertainties related to population forecasting.

In this paper, we have used such Bayesian belief network models, where probabilities have been learned directly from the observed data, often called machine learning (Mitchell 1996). This kind of Bayesian statistics, applied here in predictive stock assessments, might address some of the problems in traditional assessment models, where expert knowledge may be needed in order to make any assessment or prediction of the stock, as this new model is entirely data-driven. However, some interactions between species might arise in the model with indirect, outside factors affecting the species similarly. These hidden variables, e.g. temperature, fishing mortality and food scarcity, may cause stock to fluctuate in a similar manner, causing correlations between them even though they do not interact directly with each other. However, this kind of information is very useful when one species is predicted based on estimates available from another species, as we may not know what causes the

variation and correlation of those species, but we may, to a certain degree of certainty, estimate by observational data, how one species will react based on what we know of another species.

Bayesian statistics is based on probability interpretation and conditional probability of an event using evidence from other variables in the estimated model. These evidential probabilities are called degrees of belief, which is similar to the odds of an event happening. As an example, we could use the coronavirus PCR-test reliability based on real world data. We use a test accuracy of 95% (THL 2021), test sensitivity of 80% (Duodecim 2020) and a prior probability of 0.44%, based on the frequency of the disease in the population, i.e. percent of the population with active coronavirus in Finland as of 22.03.2021, approximately 24000 active infections (Worldometer 2021). With this information, we can calculate the conditional probability of you having the disease if you test positive for the virus with the Bayes' theorem (Figure 1). When calculated, the equation is P(A|B) = (0,8 * 0,0044) / (0,8 * 0,0044 + 0,05 * (1-0,0044)), which gives us the answer 0,066, meaning that there is a posterior probability of 6,6% that you have the coronavirus if you test positive. Note however that the prior probability plays a huge role in the outcome of the posterior probability. If we use the same equation with a prior probability of 6,6% (the posterior probability of the first test), we get a posterior probability of 53% for the second test. As more information is gathered on other variables, e.g. symptoms, being exposed to the virus etc., the prior probability is updated based on the new information, producing a more reliable posterior probability which includes the information obtained from observational data. It is likely, that the prior probability for individuals coming to the test are not the same as in the overall population, as they usually have a reason to go to a test, e.g. symptoms or being exposed to the virus. All these conditions for the event provide the probability for the event to occur.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

**Figure 1. Bayes' theorem equation. A and B are separate events. P(A|B) is the probability of event A given that event B is true and vice versa for P(B|A). P(A) and P(B) are the prior probabilities of the events, i.e. the probability of the event without any conditions.**

While statistical analyses of stock-recruitment relationships have been abundant in fisheries research (Gilbert 1997, Myers 2001 etc.), with functions as Beverton-Holt or the Ricker functions being in common use (Ogle 2018), the use of Bayesian statistics and networks has been slow to be implemented into the research field. While there were 226 research papers published in 2010 where Bayesian networks were implemented, only one was in the field of fisheries science. Of the 1375 papers published between 1990-2010 with Bayesian networks as a focus point, only 4,2 percent was in the field of environmental sciences (Aguilera et al 2011), even though Bayesian networks is a useful modeling tool that can deal with uncertainty (Fenton & Neil 2011), which is a major challenge in both fisheries research (Patterson et al 2001) and management (Fulton et al 2011). Papers from the fields of computer sciences and mathematics were the most dominating with 27.3 and 20.9 percent respectively of the published Bayesian networks papers. Of the 128 papers that studied environmental sciences with Bayesian networks, Environmental science and ecology and Water resources were the most dominant (37 and 30 papers, respectively).

Bayesian analysis methods such as Bayesian networks (Beygelzimer et al 2014) and Markov chain Monte Carlo (MCMC) (Brooks 1998) are often overlooked as a complementary or primary analysis for stock assessments. Bayesian networks are graphical models that represent a set of variables and the direct or indirect relationships between these variables. The indirect relationships are called conditional dependencies and are, in short, the relationship between two events through a third event, but having no interaction without the third event.
MCMC is a method that combines Monte Carlo methods, i.e. methods that randomly sample a probability distribution, and Markov chain, a method that gives probabilities

of each possible event based solely on the state of the previous event. MCMC is explained by Gasparini in the 1997 book "Markov chain Monte Carlo in practice" as "…essentially Monte Carlo integration using Markov chains. […] Monte Carlo integration draws samples from the required distribution, and then forms sample averages to approximate expectations. Markov chain Monte Carlo draws these samples by running a cleverly constructed Markov chain for a long time".

We suggest that Bayesian analysis of stock assessments with e.g. Bayesian networks and conditional probabilities should be used for building and analyzing a model for interspecies correlation in recruitment and stock analysis. Such models can then be used to provide the first estimates of a year-class, before anything else is known, or as a prior estimate, to be updated e.g. with surveys or catch data once they are obtained, later on (Figure 2). The model could potentially be used to learn from stocks that are geographically close or even different, but biologically relatively similar species. This could be achieved by analyzing historical stock recruitment and year-class fluctuations and making predictions with an associative model by utilizing correlations, which can be described by conditional probabilities utilized in Bayesian networks.
This could be potentially a very cost-effective way of doing stock and recruitment assessments and recruiting year-class predictions, as stocks or species could be used as "surveys" for each other to produce posterior probabilities. Cost-effectiveness is key in this method, as the surveys conducted currently are costly and labor intensive.

This kind of modelling needs, however, stock estimates and estimates from a long time period, i.e. time series, so that reliable conclusions can be made from the observed data (or model-based, estimated stock parameters) describing fisheries and survey-based estimates year-classes. For this purpose, we used the RAM Legacy Stock Assessment Database (referred here as "RAM") (Ricard et al, 2011), which was developed to further continue the work started by the Myers Stock-Recruitment Database (Myers et al, 1995). Using the data provided in RAM, we will obtain a year-class estimate for the conditional probability analysis.

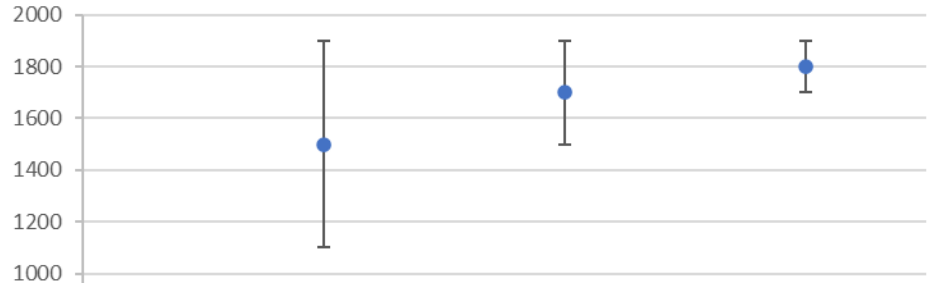


**Figure 2. Example of estimation error decreasing with more information provided (non-data based example). X = Estimation confidence using e.g. catch data, catch and acoustic surveying data and catch, acoustic surveying and Bayesian network data. Y = SSB**

The use of recruitment data in models, which are used for predicting or estimating fluctuations in stocks, could potentially decrease the uncertainty of the year-class predictions. Predicting recruitment on top of the mature stock is necessary for future predictions needed e.g. in TAC- based policy. The significance of this information increases the more the stock is exploited, i.e. when number of year-classes is low in stock and in catches.

While the concept of recruitment is not always precisely defined, as it can be the recruitment to catchable size, or the number of 0-year old fish, recruitment is still a broadly used concept in fish stock assessment models. If the same definition of recruitment is used across studies, the results are potentially comparable.

Conditional probabilities provided for the stock predictions are not used in the same way as in Bayesian networks, i.e. using all stock estimates available to determine the year-class size of the stock of interest. Figure 3Figure 3 illustrates the dependencies between two North Atlantic stocks, haddock and capelin, using linear regression. It illustrates that the uncertainties are high if only one species is used to predict the target species, but the correlation, even though low, can potentially be used to cut away some of the high uncertainties.

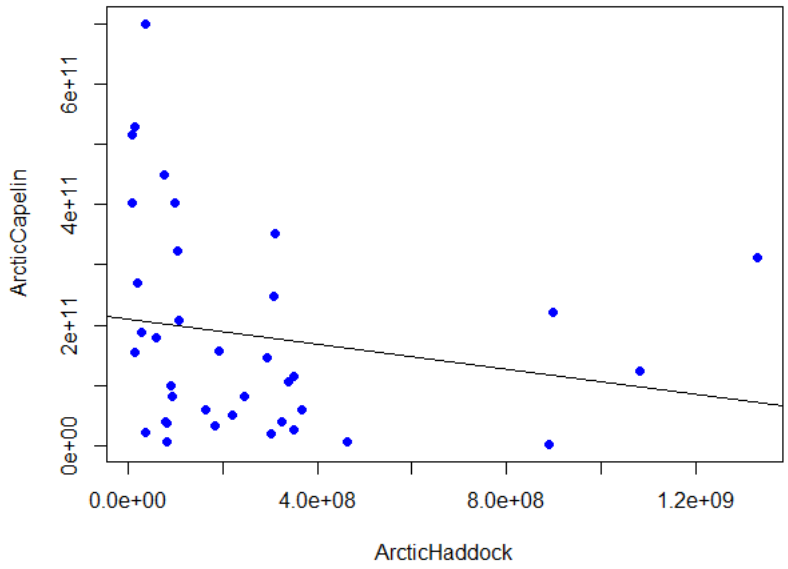


**Figure 3. Linear regression of haddock and capelin recruitment, number of individuals**

The primary aim of this thesis is to examine the possibility of modeling interdependencies between fish species and thereby decrease uncertainties in a cost-effective way, as an alternative to more costly expansion of surveys, which is a standard way to decrease uncertainties in stock predictions. We assess whether our technique would be useful in the improvement of stock recruitment and biomass assessments by lowering uncertainty and standard deviation of key stock estimates. We hope to produce a cost-effective forecasting and prediction method that would utilize natural interactions between different species and species and hidden variables. Successfully implemented, this approach would be a useful example of utilizing ecosystem knowledge in operational stock assessment, needed especially in TAC based management.

## MATERIALS AND METHODS

The data used in this analysis is based on the data available in RAM (downloaded 03.01.2019) and processed with statistical software “R” and the open source environment R-Studio. Bayesian structure learning algorithms were used with the “R”-package “BNLearn” (Scutari 2010). We use four species, of which three are

commercially or recreationally important fisheries in the Atlantic, Atlantic cod (*Gadus morhua*), pollock (*Pollachius virens*), haddock (*Melanogrammus aeglefinus*) and one is a commonly important foraging species, capelin (*Mallotus villosus*).

For the structure learning analysis, we used cod, haddock, capelin and pollock stock recruitment and spawning stock biomass (SSB) data from the northern Atlantic ICES subareas 1 and 2 (Northeast Arctic), excluding the capelin stock from Division 2.a west of 5 degrees West (Barents Sea capelin) (**Virhe. Viitteen lähdettä ei löytynyt.**) (ICES reports from 2016). Recruitment and SSB for each species were treated as separate variables, making a total of 8 variables. Stock identification codes used in RAM were CODNEAR, HADNEAR, POLLNEAR and CAPENOR for cod, haddock, pollock and capelin, respectively. The data used was isolated and combined into a single data frame to simplify the environment, i.e. the recruitment and SSB data of the target species for the target years were made into a single data frame. We erased any empty years so that each year had a data point for each stock, as the analysis used here require complete data sets, not being able to manage missing data. Data was standardized for the recruitment data, where the age (=recruitment year) would be the same across species, as different species have different recruitment age. For the species, recruitment ages reported were as follows: age 3: cod, haddock, pollock and age 1: capelin. Recruitment is reported as number of individuals and SSB as weight in metric tons. With the complete data set the recruitment and SSB data available will be for the years 1977-2013.

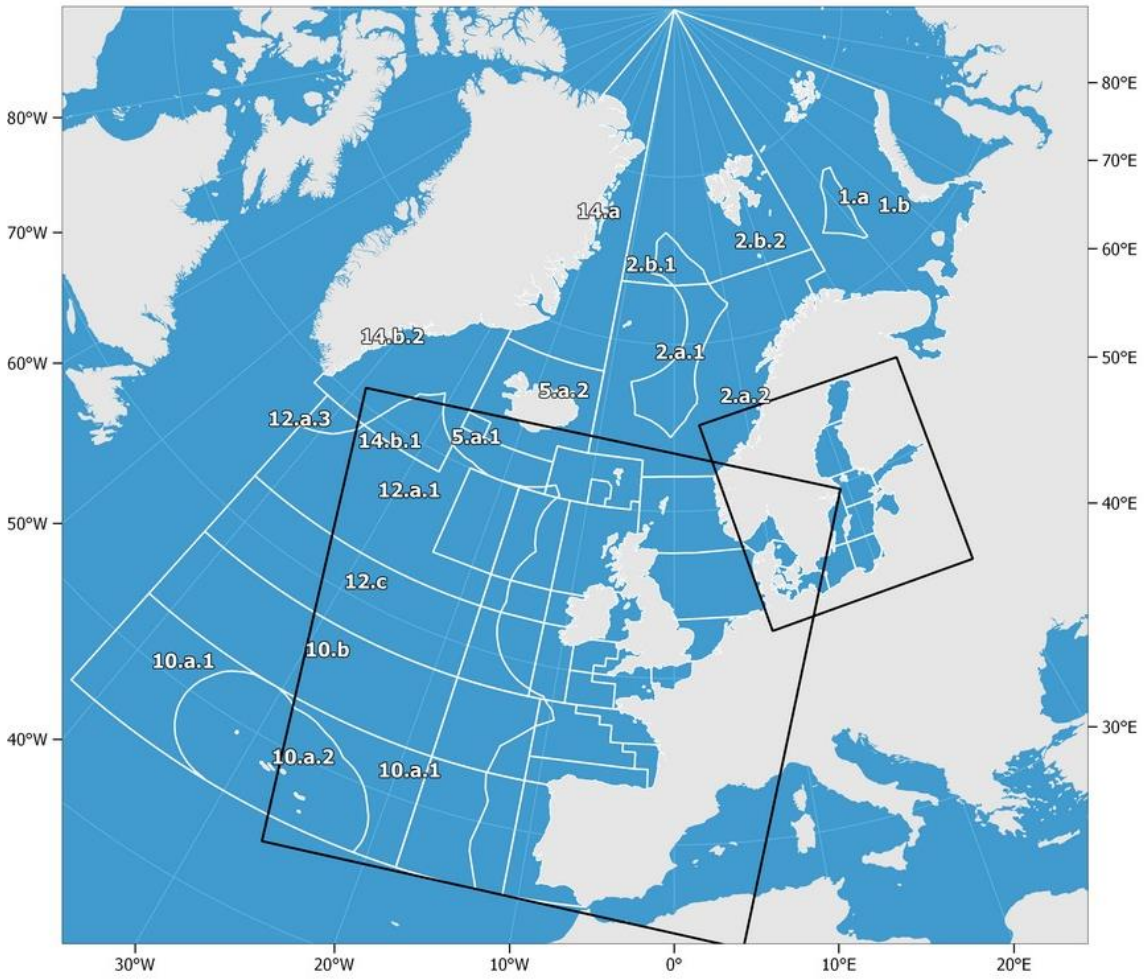


**Figure 4. Map of major fishing area 27, with subareas 1 and 2 to the top right. © FAO, 2021**

Preliminary data analysis contained a plot of the species recruitment and SSB produced with *ggplot2::ggplot* and a correlation data matrix produced with *GGally::ggpairs()* in R software. Bayesian analysis of the data went through several steps to determine the best algorithms to use. Discretization of the data was done in secondary analysis, as the initial data only had 37 data rows, with 4 species and two variables per species, with a total of 296 data points. Gaussian Bayesian structure search algorithms were primarily used, as the use of continuous data is more applicable for minor data sets (Van de Schoot & Miocević 2020). The algorithms build the network structure based on the data and gives links between species depending on the strengths of their interactions that may be quantified based on the data provided. The search algorithms used were the Tabu Search algorithm (TABU) and the Hill Climbing (HC) algorithm (Russell & Norvig 2009) for continuous data and PC (Colombo & Maathuis 2014) and Grow Shrink (GS) (Margaritis 2003) algorithms for discrete data. The difference between TABU and HC is best explained in Beretta et al 2018. As the structure of the

network was unknown, the network was allowed to be learned by the algorithm itself based solely on the observed data points, i.e. making a completely data defined model. However, the network structure learned with TABU was used for the whitelisting of the structure for the discrete data set, as the discrete data set was too small to be used for learning. Whitelisting means the use of an already learned network structure and using that structure with another Bayesian algorithm and letting the algorithm to calculate e.g. arc strengths of the given structure. Without whitelisting, the PC and GS algorithms did not find any interactions between the species. The inherent bias that whitelisting causes, i.e. giving an algorithm a ready-built structure, limits the algorithm to the given structure and will not analyze the data in a way that is natural to the algorithm itself. With larger timeseries, these methods could possibly be used as is, without whitelisting, but for this thesis we will analyze how these whitelisted, discrete methods function.

The data set was discretized with equal interval width, i.e. 1-300, 301-600 etc., with 4 factors (Low, MediumLow, MediumHigh, High) for each variable separately using *arules::discretizeDF()* (**Virhe. Viitteen lähdettä ei löytynyt.**).

**Table 1. The intervals for discretization, factors Low-MediumLow-MediumHigh-High, discretized using equal interval classification method**

| ArcticCodR | ArcticCodR | ArcticHaddockR | ArcticHaddockR | ArcticCapelinR | ArcticCapelinR | ArcticPollockR | ArcticPollockR |
|---|---|---|---|---|---|---|---|
| Low | [1.3e+08,4.5e+08) | Low | [9.48e+06,3.4e+08) | Low | [2e+09,1.76e+11) | Low | [7.18e+07,1.55e+08) |
| MediumLow | [4.5e+08,7.7e+08) | MediumLow | [3.4e+08,6.7e+08) | MediumLow | [1.76e+11,3.51e+11) | MediumLow | [1.55e+08,2.37e+08) |
| MediumHigh | [7.7e+08,1.09e+09) | MediumHigh | [6.7e+08,1e+09) | MediumHigh | [3.51e+11,5.26e+11) | MediumHigh | [2.37e+08,3.2e+08) |
| High | [1.09e+09,1.41e+09] | High | [1e+09,1.33e+09] | High | [5.26e+11,7e+11] | High | [3.2e+08,4.03e+08] |
| | | | | | | | |
| ArcticCodSSB | ArcticCodSSB | ArcticHaddockSS | ArcticHaddockSSB | ArcticCapelinSS | ArcticCapelinSSB | ArcticPollockSS | ArcticPollockSSB |
| Low | [1.03e+05,7.42e+05) | Low | [4.97e+04,2e+05) | Low | [1.7e+04,9.9e+05) | Low | [7.22e+04,2.05e+05) |
| MediumLow | [7.42e+05,1.38e+06) | MediumLow | [2e+05,3.5e+05) | MediumLow | [9.9e+05,1.96e+06) | MediumLow | [2.05e+05,3.38e+05) |
| MediumHigh | [1.38e+06,2.02e+06) | MediumHigh | [3.5e+05,5.01e+05) | MediumHigh | [1.96e+06,2.94e+06) | MediumHigh | [3.38e+05,4.7e+05) |
| High | [2.02e+06,2.66e+06] | High | [5.01e+05,6.51e+05] | High | [2.94e+06,3.91e+06] | High | [4.7e+05,6.03e+05] |

After the model was learned, the network was visualized using *bnviewer::viewer()* and scores for the networks themselves and for singular nodes were calculated with *score()*. Network scores are used to test the models so we can choose the one with the best fit (Portet 2020). For TABU and HC, the scores were computed using multivariate Gaussian Akaike Information Criterion (aic-g) (Akaike 1998), Bayesian Information Criterion (bic-g) (Schwarz 1978) and log-likelihood (loglik-g)(Koller & Friedman 2009) scores (Geiger & Heckerman 1994). For PC and GS, the Akaike Information Criterion

(aic), Bayesian Information Criterion (bic) and log-likelihood scores for discrete networks were used (Chickering 2013), for comparison. Several score metrics are used to estimate the performance of the different models while also testing the usability of the scores. Different score metrics have different assumptions that may or may not result in the metrics favoring different models (Yang & Chang 2002). In short, AIC is better when false negatives have a stronger significance in the model and reconstructing slightly denser networks, while BIC is better when false positives have stronger significance and being less affected by overfitting a model (Beretta et al 2018). BIC also favors smaller, non-complex models, while loglik favors large, more complex models (Koller & Friedman 2009).
The scores of the networks are arbitrary, meaning that they can only be used to compare the quality of one model relative to the other models in the same category (TABU and HC, PC and GS) (Portet 2020) and does not tell us the absolute quality of the model.

The strength of the arcs were computed with *arc.strength()*, the criterion being the score functions aic-g, bic-g and loglik-g for continuous variables and aic, bic and loglik for discrete variables (Appendix 1). The strength of the arc measures the gain or loss of the score for the network if the arc would be removed. It also indicates the maximum effect that a parent node can have on a child node (Oniśko & Druzdzel 2014). Negative values show a stronger relationship and would reduce the network score if the arc would be removed.

The parameters of the structure were fitted with *bn.fit()*, with Maximum Likelihood parameter estimation as the method. Fitting a model is in short making a generalization of the model so that forecasting and predictions are possible. The standard deviations of residuals were calculated with *sigma()* to determine the fit of the model per variable.

Cross-validation of the models were also implemented, to determine the mean loss of the model predictions (Koller & Friedman 2009). A *k-fold*-method was used with *bn.cv()* with 100 runs per cross-validation, with data split into 9 of equal size.

New data available in RAM in 2021 (downloaded 15.02.2021) was used as evidence and as the comparison for the predictions of the models based on data available in 2019 in blind tests to test the prediction accuracy. This new data from RAM was not loaded into the "R" environment, but rather manually extracted from the Excel-file that is included in the data folder.

The fitted structure was used for predicting the existing values of haddock and cod SSB and pollock recruitment (Figures 14, 15 and 16) with *predict()* and the predictions were plotted using *ggplot()*. The stochastic *bayes-lw*-method was used for predictions, as the *parents*-method would generally give a mean value of the total timeseries of the variable for every observation of the that variable. This is called a flat-line or no-trend model (Jewson & Penzer 2004). However, as a stochastic method, the predicted values will change between runs, giving the method a somewhat unrepeatable, but also dynamic nature, which has been implemented in models used in real-time predictions (Corman & Kecman 2018). The changes in predictions are caused by simulation noise produced by the method, i.e. variation in the simulated predictions caused by e.g. differing emphasis on coefficients. The predictions were made to demonstrate the behavior and fitness of the model on different variables, while also not being an example of how well or poorly the model can predict.
To see how well the structure may be learned with less data, we made networks with a training set of the first 31 years from the original data and predicted the last 5 years of each variable by using the last 5 years of the other variables, i.e. the test set, as evidence, also produced with the *predict()* algorithm and *bayes-lw*-method (Table 5) The mean error, root mean squared error, mean absolute error, mean percentage error and mean absolute percentage error were calculated using the *accuracy()*-command in the *forecast*-package. Mean average percentage error shows the actual error percentage between the predicted values and the actual variables, but not if the prediction is over or under the actual value. The mean percentage error shows if variables are mostly under-or overpredicted. A negative MPE means that the actual values are on average the given percentage below the predicted value and vice versa.

The fitted structure was also used for conditional probability queries using *cpquery()*. It gives the probability of an event, in our case the probability of a certain SSB or recruitment of a species, on the condition of the evidence given, i.e. a certain SSB or recruitment of one or more other species in the available data set (Koller & Friedman 2009). This conditional probability distribution may be used as a prior for future Bayesian stock predictions or as a calculated probability of a stock assessment based on traditional assessment models. If the traditional assessment model gives a SSB or recruitment value of X, we may use the Bayesian model to calculate the probability of the variable being value X based on the model. Queries are most useful for discrete variables, as there is a finite number of possible events that the method calculates the probability distribution for, which in our case is four discrete events.

We also generated random conditional probability event samples using *cpdist()*, with logic sampling as the method (Yuan & Druzdel 2006) and 10000 iterations used (n = 10000) per test to maximize efficiency, i.e. speed versus accuracy. *cpdist()* gives a random number of samples of an event, i.e. recruitment or SSB of the target species, on the condition of evidence given, i.e. recruitment or SSB values of other species. The mean value of the random event samples was calculated, and the prediction error was calculated based on the actual value of the new data available on RAM database. As an example, if the generated event samples for Cod SSB would be 200, 400 and 600, then the mean, i.e. 400, would be used as the value which would be compared to the actual Cod SSB for that year found on the RAM database.

The R-package *cpdist()* was used for each node of the TABU-network with the parental nodes used as evidence (Appendix 1), with the deviation from the actual value of the evidence ranging between approximately 0.5 to 3% below and above the event (Table 6). This was made so that the model would produce event samples more concisely. Using only the exact value for the evidence (e.g. evidence: cod SSB == 2.47e+06) would mostly produce no event samples even with higher number of iterations (tested n = 100.000 and n = 1.000.000).

As *cpquery()* and *cpdist()* are based on Monte Carlo particle filters (Chen 2003), the results will vary between runs and may sometimes not give a value for the event, i.e. target species SSB or recruitment, due to simulation noise. This is why a single parent

was used for the predictions in *cpdist()*, as for reasons unknown *cpdist()* would not produce an event for pollock SSB with all parents (haddock recruitment and SSB and cod SSB) as evidence.

## RESULTS

The results show that there is a relatively good prediction possibility for the continuous variables, with an error ranging between 1 to 40 percentage error between the actual value and the predicted value, i.e. the percentual difference between the actual value and the predicted value. In contrast, the categorical variable conditional probabilities had poor success, with only 1 out of 5 variables having the highest probability for the category that was the actual value. The predictions with the training set-test set-setting were poor, as the training set did not have values to train with that were present in the test set. These results mean, that there is to some degree either direct or indirect interactions between these species, e.g. predation or similar responses to environmental factors.

Cod and haddock SSB had a significant rise from the year 2000, which may cause the correlations to be so high. These were also connected in the Bayes network, had the strongest arc strength of both the TABU and HC networks and had good prediction accuracy for both continuous and categorical analysis.

## Explanatory analysis

The variation of the recruitment and SSB was highest for capelin. With pollock and cod, recruitment was fairly stable for the whole timeseries. Haddock recruitment had a rise between 1980-1983, after which it has had a somewhat stable fluctuation and has had two larger spikes since, in 1990 and 2005 (Figure 5). The SSB for cod and haddock has been on an increasing trend since 2000. Pollock biomass had an increasing trend since 1987, with a downward trend since 2005 (Figure 6).


**Figure 5. Species recruitment in number of individuals for years 1977 to 2013. Note the logarithmic scale**

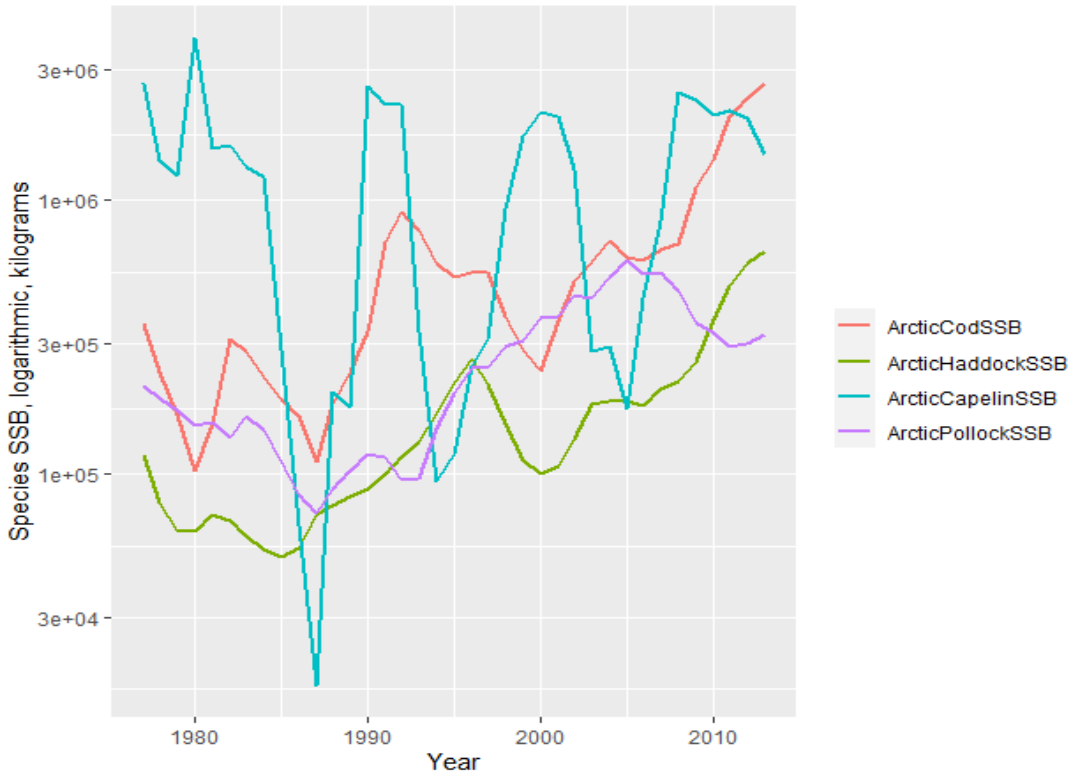


**Figure 6. Species SSB in metric tons for the years 1977-2013. Note the logarithmic scale**

Pearson's linear correlation coefficient produced a strong positive correlation between haddock SSB and cod SSB (0.96), haddock recruitment and cod recruitment (0.72) and moderate positive correlation between haddock recruitment and pollock SSB (0.56) and a weak positive correlation between pollock SSB and cod recruitment (0.34) and pollock SSB and haddock SSB (0.36). Pollock recruitment and capelin SSB and recruitment were variables with only very weak correlations (<0.3) with other variables (Figure 7). The statistical significance (p-value) has no role here, as the question is only about flow of information between the variables.

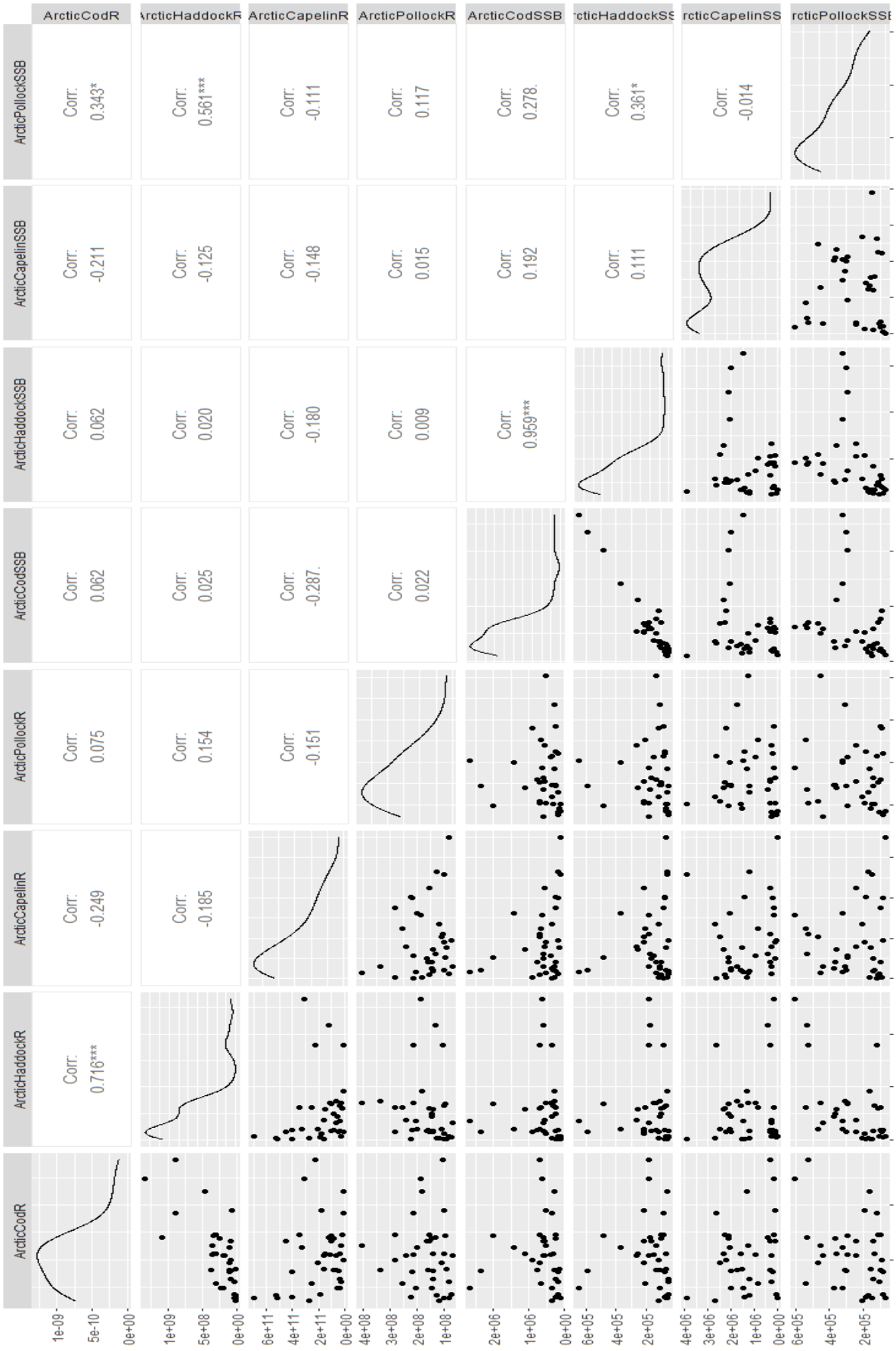


**Figure 7. Pearson's linear correlation matrix between fish species recruitment and SSB, produced with ggpairs(). Scatterplots between pairs on the left are the mirrors of the correlations on the right.**

## Gaussian, score-based Bayes networks

### TABU and Hill-climbing-search algorithms

TABU and HC-search algorithms had similar structures, with the differences being that capelin recruitment was connected to cod SSB by TABU and to haddock SSB by HC and that arc direction was reverse for cod recruitment and haddock recruitment and for cod SSB and haddock SSB. Both search algorithms found the strongest arcs between cod SSB and haddock SSB (AIC strength score of -47,41 for both) and moderate arc strengths for cod recruitment and haddock recruitment and for haddock recruitment and pollock SSB (AIC strength score of -12,27 and -7,94, respectively, Figures 7 and 9). The strongest arcs were the same species pairs that had higher correlation coefficients (Figures 6 and 8). Pollock recruitment and capelin SSB were left out of the networks by the algorithms most likely due to low interactions with the other species. All arc strengths for the arcs in the TABU and HC networks are available in Appendix 1.

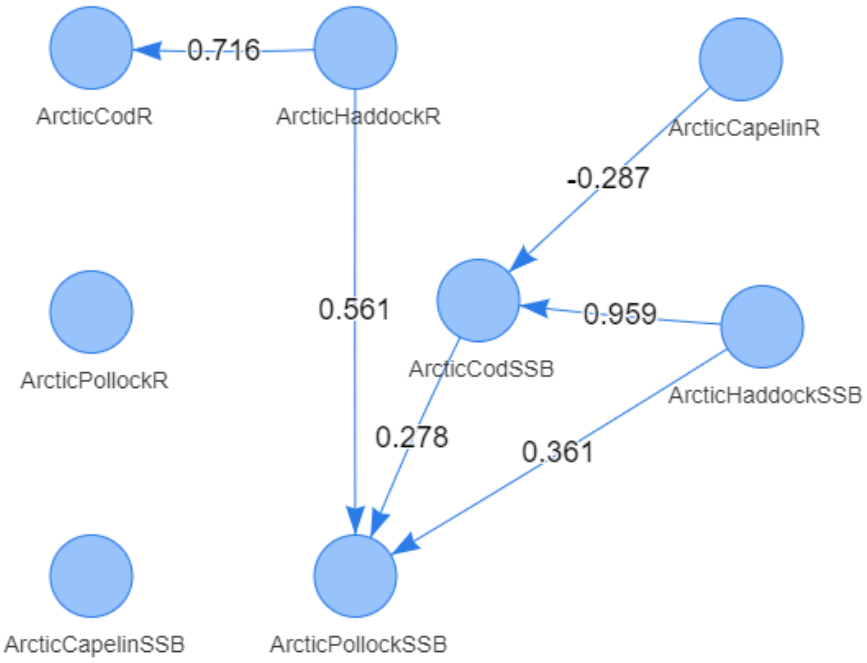


**Figure 8. TABU-search algorithm Bayes network with cod, pollock, capelin and haddock. R = recruitment, SSB = spawning stock biomass. Pearson's correlation coefficient shown between connected nodes. Visualized with bnviewer::viewer()**

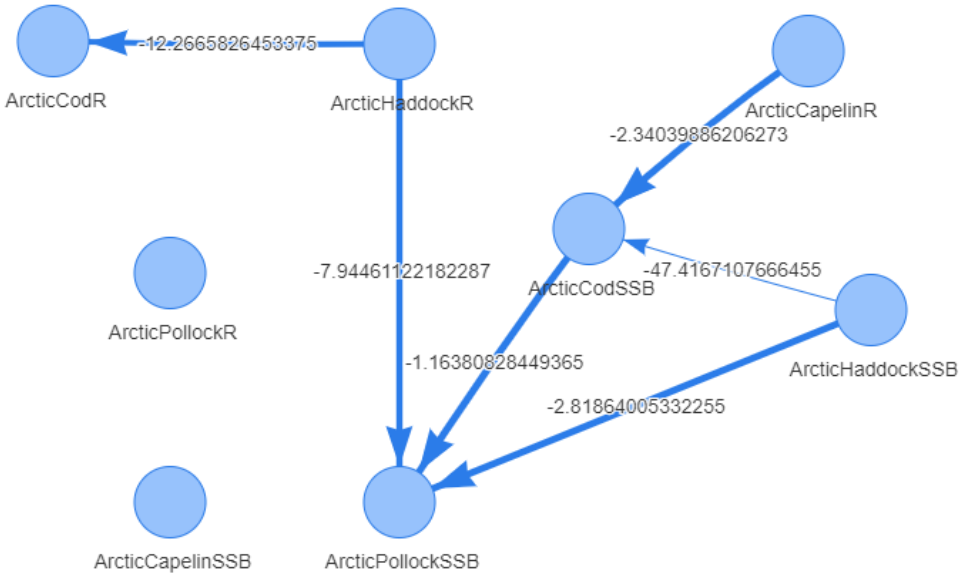


**Figure 9. Arc strength computed with Gaussian Akaike information criterion (AIC-g) for the TABU-search algorithm. Visualized with bnviewer::strength.viewer(). Arc thickness tells if the direction of the arc is strong or if the direction could be either way. A negative strength value means a stroger link and removing that link would decrease the overall score (prediction capacity) of the network.**

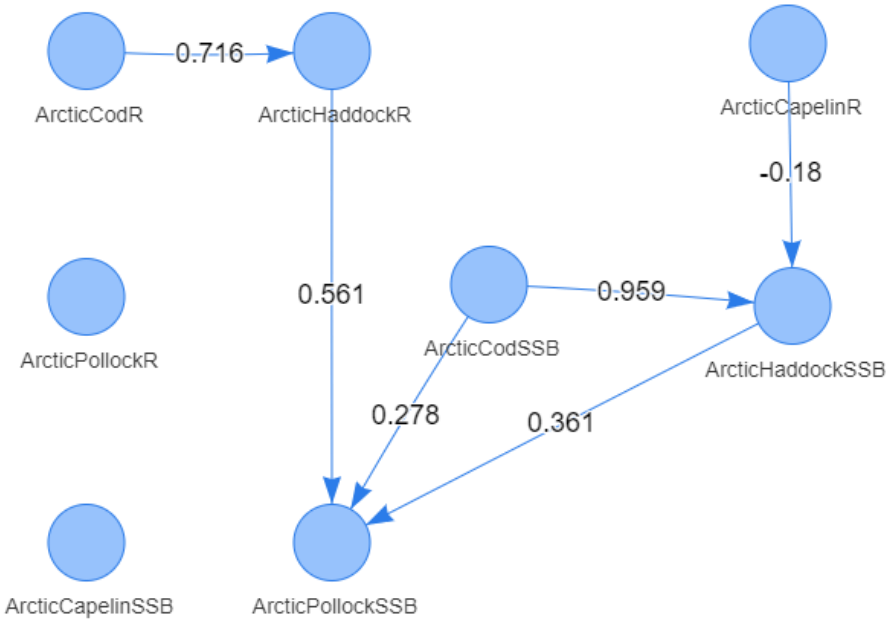


**Figure 10. Hill-climbing (HC)-search algorithm, cod, pollock, caplin and haddock. Correlation coefficient shown.**

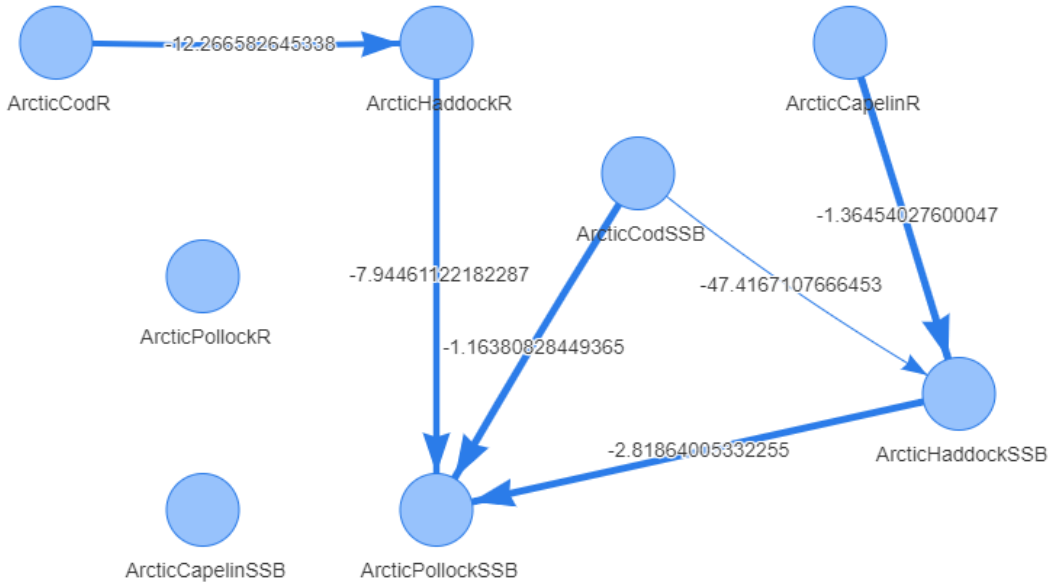


**Figure 11. Arc strength computed with Gaussian Akaike information criterion (AIC-g) for the HC-search algorithm.**

## Score and cross validation

The scores of the nodes were similar for both TABU and HC with loglik, AIC and BIC, with marginal difference in overall scores across all score functions (Table 2). Higher negative values indicate a better fitting of the model, i.e. the least information loss, which means that HC has a marginally better fit. Note that the scores are arbitrary and are to be compared to the same variable of the same score function across models for easily comparable results.

The mean log-likelihood loss of information calculated with k-fold cross-validation was 144.1984 for TABU and 144.2432 for HC. A marginally higher outlier was identified with TABU (Figure 10), but the loss was essentially the same for both algorithms. The residuals of the fitted models for TABU and HC indicate that the best suited and more stable approach for predicting cod, haddock and pollock SSB (Table 3) would be the TABU-algorithm.

**Table 2. Score for the algorithms (TABU, HC) per score function (loglik, aic, bic), including scores per variable and combined score. Higher negative score for the same score function between algorithms indicates a better fit of the model. Note that the scores are only comparable between the algorithms of the same variable and score function (i.e. Cod R TABU aic -> Cod R HC aic)**

| | TABU loglik | HC loglik | TABU aic | HC aic | TABU bic | HC bic |
|---|---|---|---|---|---|---|
| ArcticCodR | -761,8923 | -775,1589 | -764,8923 | -777,1589 | -767,3087 | -778,7698 |
| ArcticHaddockR | -775,5463 | -762,2797 | -777,5463 | -765,2797 | -779,1572 | -767,6961 |
| ArcticCapelinR | -1009,6863 | -1009,6863 | -1011,6863 | -1011,6863 | -1013,2972 | -1013,2972 |
| ArcticPollockR | -723,2953 | -723,2953 | -725,2953 | -725,2953 | -726,9062 | -726,9062 |
| ArcticCodSSB | -494,1810 | -544,1638 | -498,1810 | -546,1638 | -501,4029 | -547,7748 |
| ArcticHaddockSSB | -491,2302 | -442,2233 | -493,2302 | -446,2233 | -494,8411 | -449,4451 |
| ArcticCapelinSSB | -562,5338 | -562,5338 | -564,5338 | -564,5338 | -566,1447 | -566,1447 |
| ArcticPollockSSB | -480,8795 | -480,8795 | -485,8795 | -485,8795 | -489,9068 | -489,9068 |
| TOTAL | -5299,2450 | -5300,2210 | -5321,2450 | -5322,2210 | -5338,9650 | -5339,9410 |

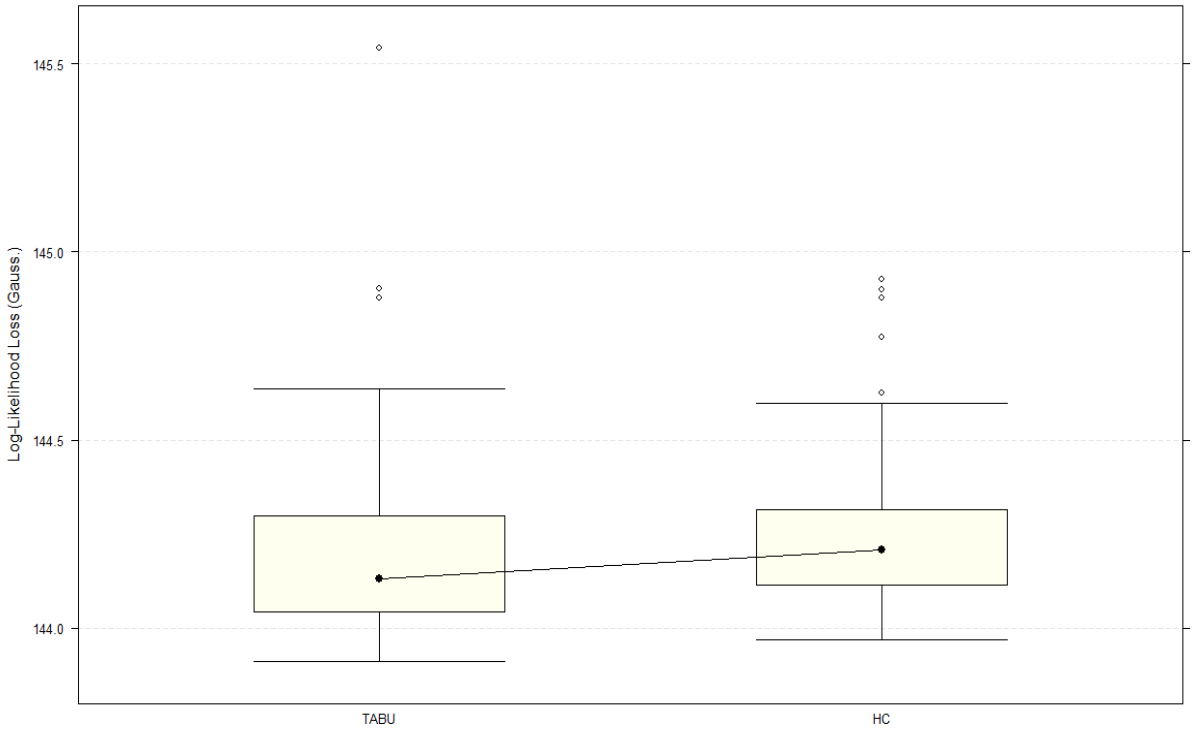


**Figure 12. Loss of data or prediction error calculated with cross-validation for TABU and HC algorithms with log-likelihood loss estimation. k = 9, runs = 100. Lower value indicates a better predicting model.**

**Table 3. Standard deviations of residuals TABU and HC algorithms for each variable**

| Residual SD | Tabu | HC |
|---|---|---|
| ArcticCodR | 217949153,0 | 307755170,0 |
| ArcticHaddockR | 310994559,0 | 220243255,0 |
| ArcticCapelinR | 174186741806,0 | 174186741806,0 |
| ArcticPollockR | 75761216,0 | 75761216,0 |
| ArcticCodSSB | 159188,8 | 598215,3 |
| ArcticHaddockSSB | 143066,9 | 39088,4 |
| ArcticCapelinSSB | 982829,8 | 982829,8 |
| ArcticPollockSSB | 112629,4 | 112629,4 |

## Discrete, constraint-based Bayes networks

### PC and GS-search algorithms

The PC and GS network structures and scores are identical to each other, most likely due to the whitelisting of the structure (Figure 11, 12).

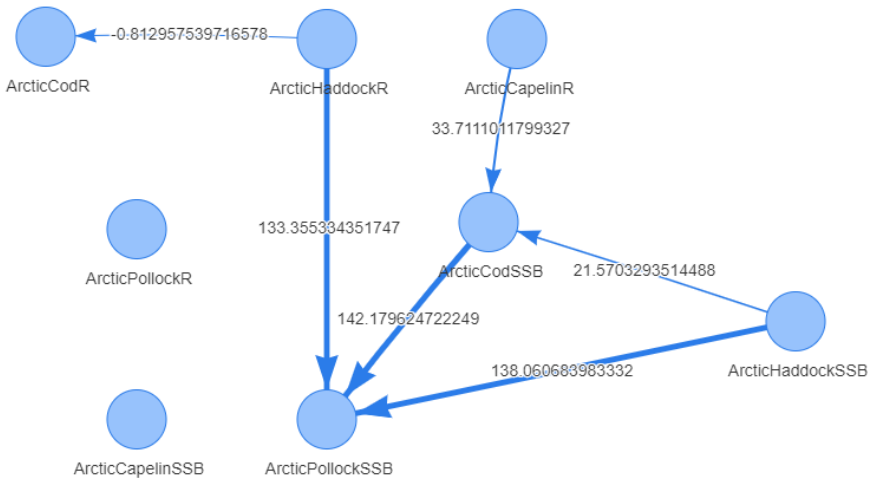


**Figure 13. Arc strength computed with Akaike information criterion (AIC) for the discrete PC-search algorithm. Network structure whitelisted or built with the structure given by the TABU-search algorithm**

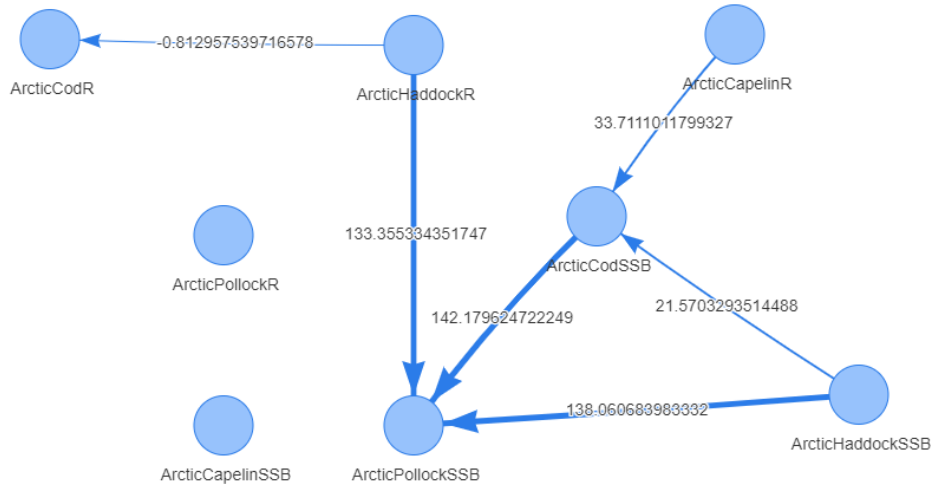


**Figure 14. Arc strength computed with Akaike information criterion (AIC) for the discrete GS-search algorithm. Network structure whitelisted or built with the structure built by the TABU-search algorithm.**

## Score and cross validation

There was no difference in scores between the PC and GS algorithms, most likely due to the whitelisting of the network (Table 4). The mean loss of data between the algorithms in the cross-validation analysis was the same at approximately 6,83 (Figure 13). Note that of the 100 cross-validation runs, the cross-validation method could not produce a loss of data for each run, with 5 NA for PC and 9 NA for GS due to the model not recognizing interactions between variables in each run and thus dropping data on several occasions during the analysis.

**Table 4. Score for the discrete algorithms (PC, GS) per score function (loglik, aic, bic), including scores per variable and combined score. Higher negative score for the same score function between algorithms indicates a better fit of the model. Note that the scores are only comparable between the algorithms of the same variable and score function (i.e. Cod R PC aic -> Cod R GS aic)**

| | PC loglik | GS loglik | PC aic | GS aic | PC bic | GS bic |
|---|---|---|---|---|---|---|
| ArcticCodR | -30,38603489 | -30,38603489 | -42,38603489 | -42,38603489 | -52,05154237 | -52,05154237 |
| ArcticHaddockR | -29,4824582 | -29,4824582 | -32,4824582 | -32,4824582 | -34,89883507 | -34,89883507 |
| ArcticCapelinR | -39,53201254 | -39,53201254 | -42,53201254 | -42,53201254 | -44,94838941 | -44,94838941 |
| ArcticPollockR | -40,38099749 | -40,38099749 | -43,38099749 | -43,38099749 | -45,79737436 | -45,79737436 |
| ArcticCodSSB | -7,543873832 | -7,543873832 | -55,54387383 | -55,54387383 | -94,20590374 | -94,20590374 |
| ArcticHaddockSSB | -31,09322185 | -31,09322185 | -34,09322185 | -34,09322185 | -36,50959872 | -36,50959872 |
| ArcticCapelinSSB | -43,09067402 | -43,09067402 | -46,09067402 | -46,09067402 | -48,50705088 | -48,50705088 |
| ArcticPollockSSB | -23,73052098 | -23,73052098 | -215,730521 | -215,730521 | -370,3786406 | -370,3786406 |
| TOTAL | -245,2397938 | -245,2397938 | -512,2397938 | -512,2397938 | -727,2973351 | -727,2973351 |

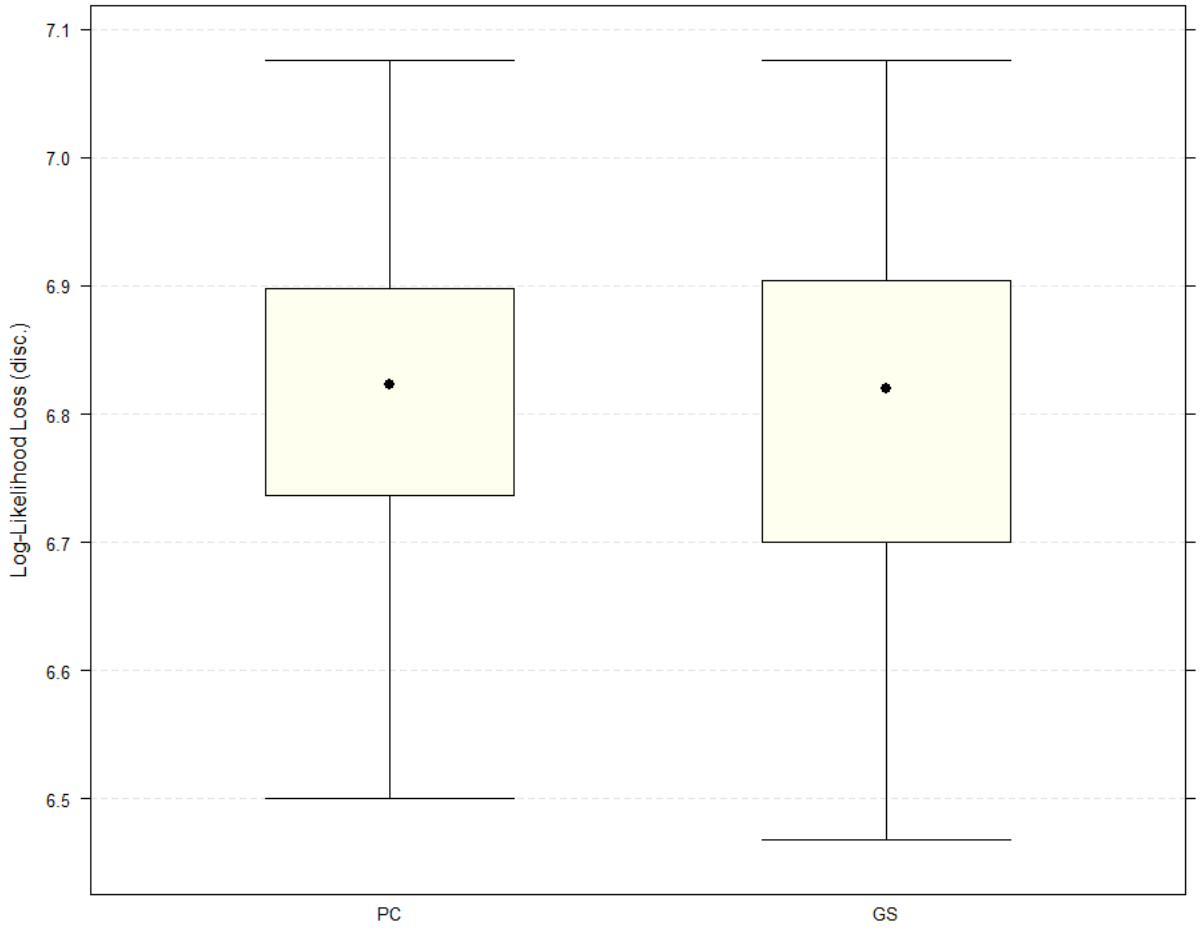


**Figure 15. Loss of data or prediction error calculated with cross-validation for PC and GS algorithms with log-likelihood loss estimation. k = 9, runs = 100. Lower value indicates a better predicting model.**

## Predictions

After the model was learned through the use of a Bayesian network algorithm, with interactions between species being in the forefront of the analysis, the predictive power of the network/models was tested. In general, the continuous models performed better than the discrete models, which is likely due to the fact that there is some loss of information on interactions in discrete model as there may only be just a few data points in a given category of a variable, e.g. only two Low SSB for cod. However, the differences in the accuracy of the predictions are unquantifiable, as the predictions produced and data they are based on are different in terms of statistical mathematics, i.e. categorical variables and predictions versus continuous.

### Predictions with continuous data

Preliminary predictions were made using training-set-test-set-setting, with all other variables in the network used to predict one variable with a training set of the first 31 rows (years) and a test set of the last 6 rows of the available data. The TABU and HC algorithms could produce moderate mean actual percentage errors (MAPE) for the predictions of the cod and haddock SSB and pollock recruitment (Table 5). These predictions had an MAPE of 29,39%, 32,77% and 36,93% for TABU and 35,70%, 23,83% and 37,83% for HC, respectively. The rest of the predictions had a MAPE of over 85%.

Cod recruitment and pollock SSB and recruitment were overpredicted by both TABU and HC, while the other variables were underpredicted.

**Table 5. Prediction outcomes of test set length 6 with TABU and HC algorithms based on a training set of length 31. ME: Mean Error, RMSE: Root Mean Squared Error, MAE: Mean Absolute Error, MPE: Mean Percentage Error, MAPE: Mean Absolute Percentage Error**

| TABU | | | | | | HC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Predicted | Actual ArcticCodSSB | | | | | Predicted | Actual ArcticCodSSB | | | |
| 625330.2 | 694000 | | | | | 621746.1 | 694000 | | | |
| 745567.6 | 1110000 | | | | | 694624.8 | 1110000 | | | |
| 948817.0 | 1410000 | | | | | 1008093.1 | 1410000 | | | |
| 1299805.6 | 2020000 | | | | | 1058984.4 | 2020000 | | | |
| 1506189.3 | 2350000 | | | | | 1066926.6 | 2350000 | | | |
| | | | | | | | | | | |
| ME | RMSE | MAE | MPE | MAPE | | ME | RMSE | MAE | MPE | MAPE |
| 491658 | 562301.1 | 491658 | 29.39891 | 29.39891 | | 626725 | 762773.2 | 626725 | 35.70207 | 35.70207 |
| | | | | | | | | | | |
| Predicted | Actual ArcticCodR | | | | | Predicted | Actual ArcticCodR | | | |
| 439321263,00 | 4.73e+08 | | | | | 449592967,00 | 4.73e+08 | | | |
| 826428336,00 | 5.54e+08 | | | | | 837765452,00 | 5.54e+08 | | | |
| 901649823,00 | 6.14e+08 | | | | | 897608852,00 | 6.14e+08 | | | |
| 1415412082,00 | 7.19e+08 | | | | | 1296289323,00 | 7.19e+08 | | | |
| 1395921194,00 | 3.99e+08 | | | | | 1472313300,00 | 3.99e+08 | | | |
| | | | | | | | | | | |
| ME | RMSE | MAE | MPE | MAPE | | ME | RMSE | MAE | MPE | MAPE |
| -443946540 | 572177513 | 457418034 | -87.12328 | 89.97138 | | -438913979 | 573893562 | 448276792 | -88.35087 | 90.33032 |
| | | | | | | | | | | |
| Predicted | Actual ArcticCapelinR | | | | | Predicted | Actual ArcticCapelinR | | | |
| 1,75721E+11 | 2.09e+11 | | | | | 1,14181E+11 | 2.09e+11 | | | |
| 28597998331 | 1.46e+11 | | | | | -21412010029 | 1.46e+11 | | | |
| -10852790029 | 3.24e+11 | | | | | -1,1199E+11 | 3.24e+11 | | | |
| -1,3968E+11 | 1.05e+11 | | | | | -3,02877E+11 | 1.05e+11 | | | |
| -1,27008E+11 | 4.00e+10 | | | | | -4,05429E+11 | 4.00e+10 | | | |
| | | | | | | | | | | |
| ME | RMSE | MAE | MPE | MAPE | | ME | RMSE | MAE | MPE | MAPE |
| 1,79444E+11 | 2.07257e+11 | 1,79444E+11 | 170.0469 | 170.0469 | | 3,10305E+11 | 3,44057E+11 | 3,10305E+11 | 359.3248 | 359.3248 |
| | | | | | | | | | | |
| Predicted | Actual ArcticCapelinSSB | | | | | Predicted | Actual ArcticCapelinSSB | | | |
| 325157.4 | 2470000 | | | | | 509110.3 | 2470000 | | | |
| 321611.7 | 2320000 | | | | | 153844.3 | 2320000 | | | |
| -424826.6 | 2050000 | | | | | -551407.5 | 2050000 | | | |
| -792739.3 | 2120000 | | | | | -1366926.8 | 2120000 | | | |
| -785696.3 | 2000000 | | | | | -2028924.4 | 2000000 | | | |
| | | | | | | | | | | |
| ME | RMSE | MAE | MPE | MAPE | | ME | RMSE | MAE | MPE | MAPE |
| 2463299 | 2488487 | 2463299 | 114.0749 | 114.0749 | | 2848861 | 2956198 | 2848861 | 133.1158 | 133.1158 |
| | | | | | | | | | | |
| Predicted | Actual ArcticHaddockSSB | | | | | Predicted | Actual ArcticHaddockSSB | | | |
| 195906.2 | 219000 | | | | | 188503.4 | 219000 | | | |
| 219431.0 | 259000 | | | | | 222366.6 | 259000 | | | |
| 228429.1 | 369000 | | | | | 268846.1 | 369000 | | | |
| 273181.9 | 488000 | | | | | 340427.7 | 488000 | | | |
| 259532.5 | 589000 | | | | | 390499.9 | 589000 | | | |
| | | | | | | | | | | |
| ME | RMSE | MAE | MPE | MAPE | | ME | RMSE | MAE | MPE | MAPE |
| 149503.8 | 187912.1 | 149503.8 | 32.77493 | 32.77493 | | 102671.3 | 121229.3 | 102671.3 | 23.8306 | 23.8306 |
| | | | | | | | | | | |
| Predicted | Actual ArcticHaddockR | | | | | Predicted | Actual ArcticHaddockR | | | |
| 320884401 | 1.07e+08 | | | | | 335663631 | 1.07e+08 | | | |
| 202529289 | 2.94e+08 | | | | | 200038594 | 2.94e+08 | | | |
| 18244498 | 1.05e+08 | | | | | 68309055 | 1.05e+08 | | | |
| -174248962 | 3.40e+08 | | | | | -90734252 | 3.40e+08 | | | |
| -372935902 | 7.92e+07 | | | | | -536205783 | 7.92e+07 | | | |
| | | | | | | | | | | |
| ME | RMSE | MAE | MPE | MAPE | | ME | RMSE | MAE | MPE | MAPE |
| 186145335 | 325735726 | 271699096 | 127.1946 | 207.1514 | | 189625751 | 354039066 | 281091203 | 151.3826 | 236.8644 |
| | | | | | | | | | | |
| Predicted | Actual ArcticPollockSSB | | | | | Predicted | Actual ArcticPollockSSB | | | |
| 336539.8 | 469000 | | | | | 336713.4 | 469000 | | | |
| 432630.5 | 360000 | | | | | 438800.2 | 360000 | | | |
| 507379.5 | 326000 | | | | | 518801.7 | 326000 | | | |
| 742786.7 | 292000 | | | | | 744577.8 | 292000 | | | |
| 851662.3 | 301000 | | | | | 850467.0 | 301000 | | | |
| | | | | | | | | | | |
| ME | RMSE | MAE | MPE | MAPE | | ME | RMSE | MAE | MPE | MAPE |
| -224599.8 | 335308 | 277583.9 | -76.97865 | 88.27589 | | -228272 | 336933.9 | 281186.6 | -78.07281 | 89.35525 |
| | | | | | | | | | | |
| Predicted | Actual ArcticPollockR | | | | | Predicted | Actual ArcticPollockR | | | |
| 172804440 | 1.09e+08 | | | | | 173372870 | 1.09E+08 | | | |
| 169735390 | 1.41e+08 | | | | | 176835822 | 1.41e+08 | | | |
| 178429408 | 2.00e+08 | | | | | 175125119 | 2.00e+08 | | | |
| 173811997 | 9.87e+07 | | | | | 172329973 | 9.87e+07 | | | |
| 171173129 | 1.44e+08 | | | | | 169411099 | 1.44e+08 | | | |
| | | | | | | | | | | |
| ME | RMSE | MAE | MPE | MAPE | | ME | RMSE | MAE | MPE | MAPE |
| -34650873 | 48460728 | 43279110 | -32.62043 | 36.93455 | | -34874977 | 49221802 | 44824929 | -32.85642 | 37.83139 |

When predicting only one year of one variable using a parent node from the learned TABU model with the actual numbers used being values gotten from RAM that were not included in the model, the predictions had good errors for cod SSB with evidence being the parent node haddock SSB and for pollock SSB with evidence being the parent node cod SSB. These had a percentage error of 8,63% and 1,28%, respectively (Table 6).

**Table 6. Predictions of variables based on the evidence of the parent node in the TABU-network. Evidence and actual values for variables based on values from 2014 available on RAM**

| | Prediction | Actual | Error (actual) | Error (%) | Evidence | Nr of samples |
|---|---|---|---|---|---|---|
| ArcticCodSSB | 2737595 | 2520000 | 217595 | 8,634722222 | ArcticHaddockSSB 6.7e+05 - 6.8e+05 | 1 |
| ArcticCodR | 490599920 | 688000000 | -197400080 | -28,69187209 | ArcticHaddockR 1.85e+08 - 1.94e+08 | 135 |
| ArcticPollockSSB | 247269,8 | 348000 | -100730,2 | -28,94545977 | ArcticHaddockR 1.85e+08 - 1.94e+08 | 115 |
| ArcticPollockSSB | 352464,7 | 348000 | 4464,7 | 1,28295977 | ArcticCodSSB 2.47e+06 to 2.57e+06 | 3 |
| ArcticPollockSSB | 206013,7 | 348000 | -141986,3 | -40,80066092 | ArcticHaddockSSB 6.7e+05 - 6.8e+05 | 1 |

Model predictions using the learned model against the whole data set the model were learned from produced well-fitting results for cod and haddock SSB, but poorly fitted results for pollock recruitment (Figures 14, 15, 16). Pollock recruitment showed a no-trend model due to no connections to the other variables in the model (Figure 8) that were used for predictions.

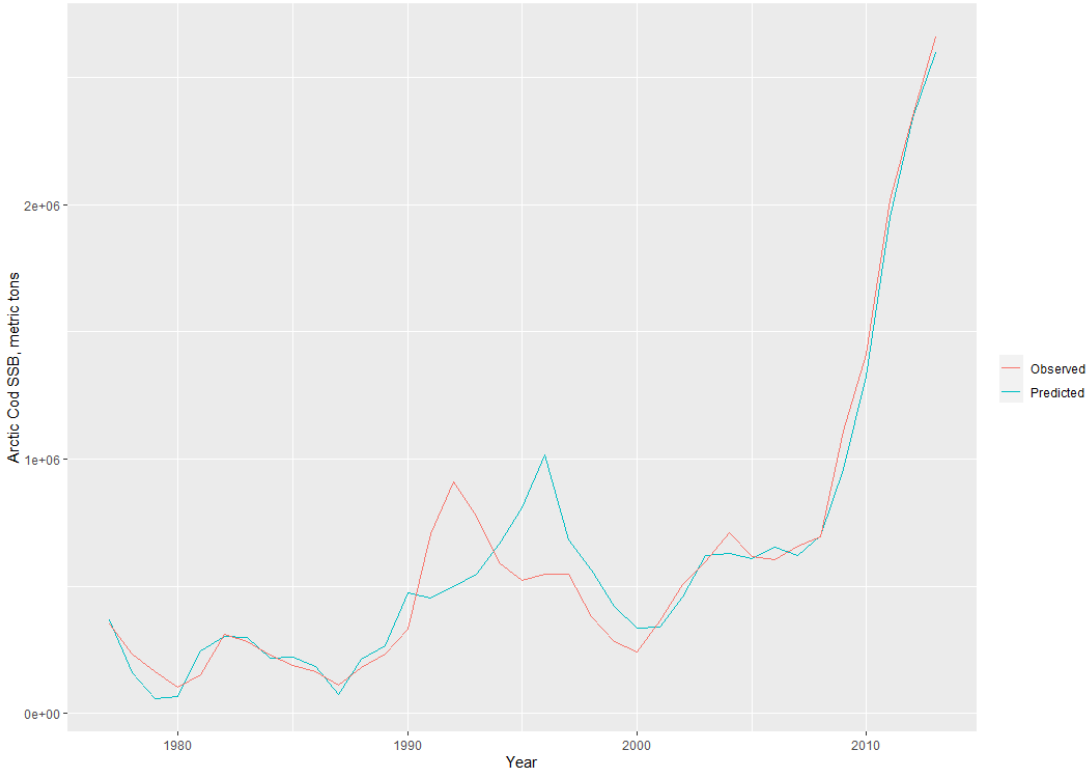


**Figure 16. Predicted values of cod SSB (blue) versus estimated values in the data set(red), predicted by using all other nodes from the fitted TABU-search model as evidence in the predict() command, with the bayes-lw-method. As a stochastic method, the absolute results vary per run**

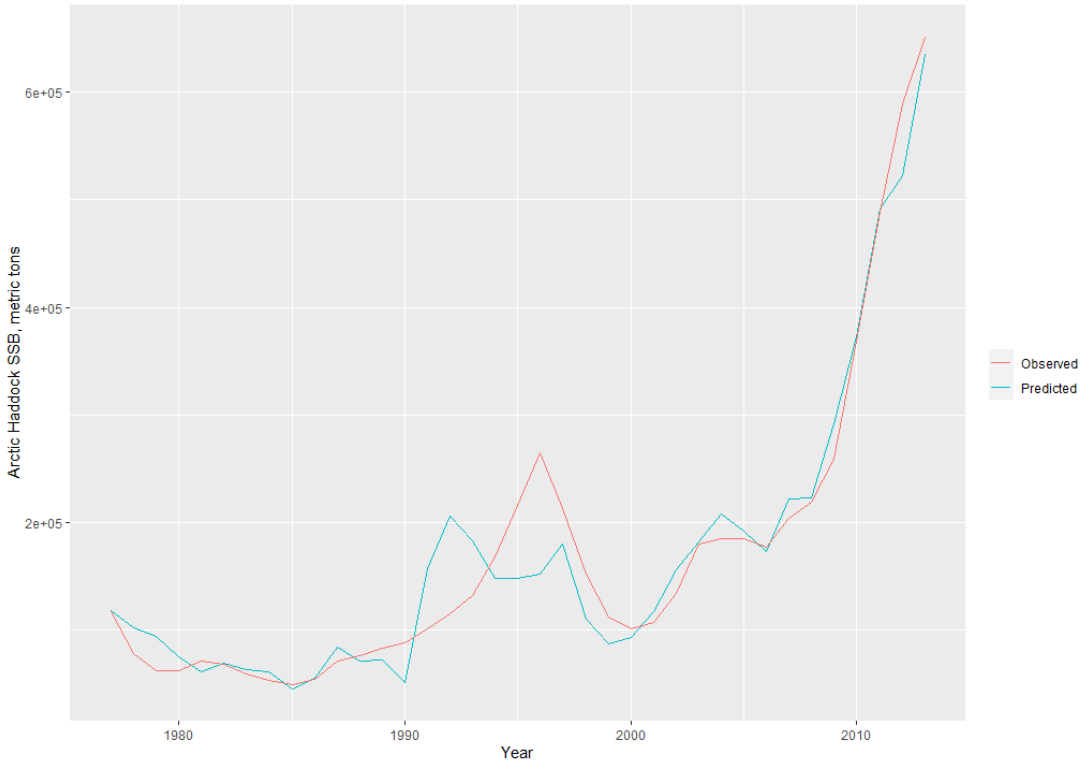


**Figure 17. Predicted values of haddock SSB (blue) versus estimated values in the data set (red), predicted by using all other nodes from the fitted TABU-search model as evidence in the predict() command, with the bayes-lw-method**

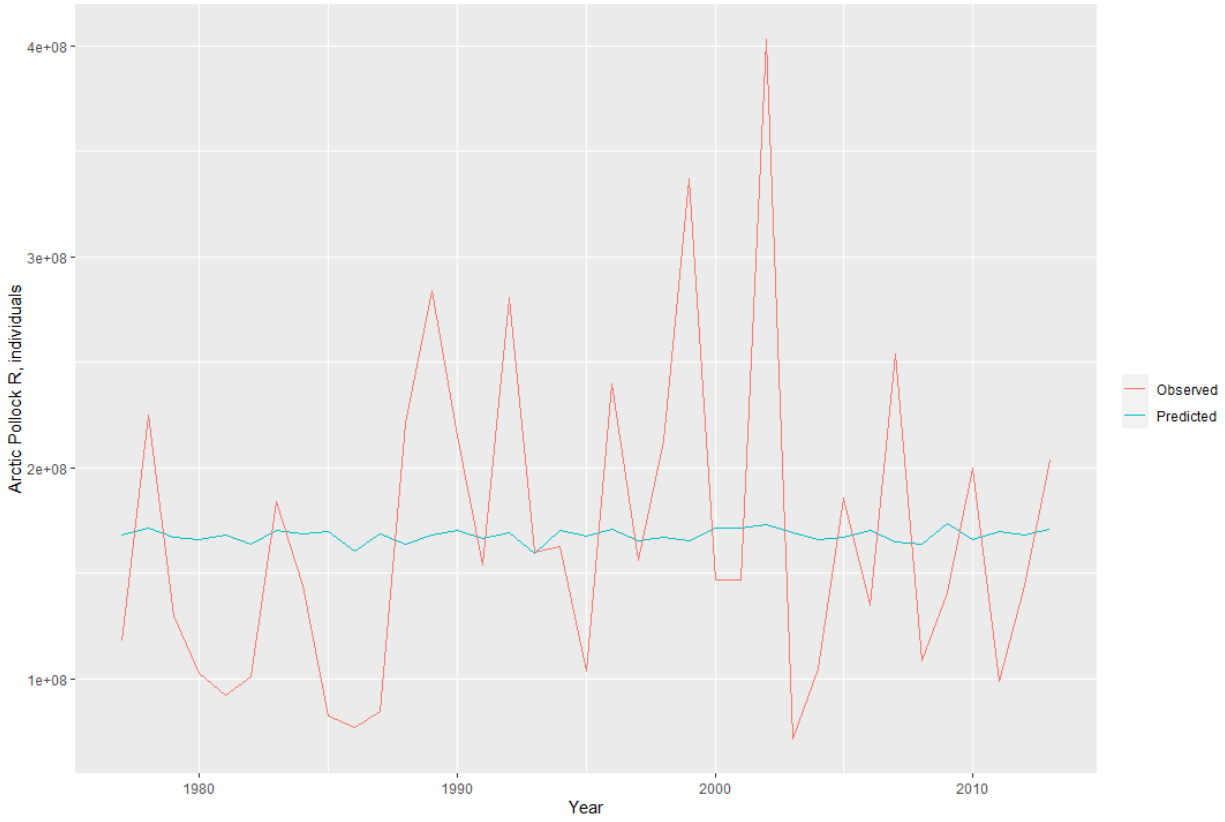


**Figure 18. Predicted values of pollock recruitment (blue) versus estimated values in the data set (red), predicted by using all other nodes from fitted TABU-search model as evidence in the predict() command, with the bayes-lw-method. A no-trend prediction can be seen (blue line), most likely due to the variable not being connected to other variables in the network.**

## Predictions with continuous data

The conditional probabilities of the variables based on the parent nodes actual value from 2014 showed a mediocre predictive probability for cod SSB and recruitment, while for pollock SSB, the conditional probabilities matched poorly with the actual value when predicted with the PC model for discrete data (Table 7).

**Table 7. Predictions of variables based on the evidence of the parent node in the PC-network. Evidence and actual values for variables based on values from 2014 available on RAM**

| | ArcticCodSSB | ArcticCodR | ArcticPollockSSB | ArcticPollockSSB | ArcticPollockSSB |
|---|---|---|---|---|---|
| Low | 0.13 | 0.526 | 0.544 | 0.12663869 | 0.108 |
| MediumLow | 0.12 | 0.458 | 0.208 | 0.64746455 | 0.662 |
| MediumHigh | 0.13 | 0.016 | 0.214 | 0.08146691 | 0.114 |
| High | 0.62 | 0.000 | 0.034 | 0.14442984 | 0.116 |
| Actual | High | MediumLow | MediumHigh | MediumHigh | MediumHigh |
| Evidence | Haddock SSB High | Haddock R Low | Haddock R Low | Cod SSB High | Haddock SSB High |

# DISCUSSION

In this analysis, I have shown that there is an obvious possibility to predict fish stock fluctuations with a data driven approach using Bayesian networks. They have an obvious potential to decrease the uncertainties of stock estimates needed in fisheries management, especially those based on biomass estimates (like TAC). Without the need for expert opinions on the correlations or causations between stocks, the models are easy to produce and with a ready-to-use-model, it is not very labor intensive to produce predictions and forecasts of stocks that are already incorporated in the data sets available.

The methodology applied found strong interactions between cod and haddock SSB, which is most likely caused by the similarities in biology between the species, and potentially due to the similarities in fishing pressure. The strong correlation also most likely caused these stocks to be somewhat more easily predictable, with a correlation of 0.96 (Figure 5) and a mean absolute percentage error for the predictions of a test set of cod SSB being 29,39% for the TABU algorithm and 35,70% for the HC algorithm and for haddock, 32,77% and 23,83%, respectively (Table 5). Correlations between cod and haddock recruitment were also strong, 0.72, however the predictions with the above method were mostly poor.

Thus, the approach presented here could be used to provide additional information for cod and haddock SSB fluctuations if used with the evidence provided from the other variables in the RAM data set. However, with respect to the high margin of error for the prediction of recruitment of these stocks and for pollock and both SSB and recruitment of capelin, the suggested methodology seems to be of smaller use for the predictions of recruitment (Table 5). However, the amount of predictions for a single year we could do for recruitment was not as comprehensive as for SSB (Table 6) due to unknown technical issues, which caused the method to not produce predictions even when the amount of iterations per prediction was raised to one million. With available computational power, this prediction took roughly 5 to 10 minutes. In future research this needs to be addressed, as recruitment is a vital point in the proposed approach.

It is however imperative that when the predictions are made, the structure of the network is analyzed and defined, and the strengths of the arcs and connections between nodes are taken into consideration. While pollock recruitment had a decent mean actual percentage error for the predictions, the predictions seem to be the mean or median of the evidence provided, as visualized in Figure 18. As the recruitment of pollock does not fluctuate as much as the recruitment of other stocks, with weak interactions with other variables, i.e. having no parental nodes, the coefficients used in the predictions give a so called flatline or no-trend prediction (Jewson & Penzer 2004). This means that the predicted value of pollock will only be the mean value of the previous years, without taking evidence given from other species into consideration. While it has a good percentual prediction error, it does not give us results that would significantly differ from the results of a trendline calculated with the mean of the values of the variable. Keeping the variable in, however, may slightly increase the prediction accuracy of the other variables due to increased quantity of data, which is often true with machine learning algorithms that do not have a pre-set prejudice towards and between variables. This means that even though the interactions between species according to the method is very low, the interaction may give us a marginal amount of information of other species based on the small trends and correlations of pollock compared to the other species.

For a small data set, in terms of machine learning, discretizing the data does not seem feasible, as the loss of knowledge in the discretization process is quite significant and would most likely be the cause of the mediocre predictions (Kotsiantis & Kanellopoulos 2006). However, if used with massive databases where data input is continuous and leads to slower analysis due to the high need of computational power, the discretization of data may be acceptable and even practical, as the data used will be divided in to a limited amount of categories or factors instead of a theoretically unlimited amount of differing numbers, e.g. category 1 representing any number between 1000 and 2000. In our case, however, the constraint-based algorithms did not see any connections between the discretized variables, which is why the implementation of whitelisting was needed, which may cause some bias and is thus not purely a data-based analysis with the algorithm provided. As the structure is given

to the algorithm, the algorithm cannot work as intended in building an structure optimized to the emphasis of the algorithm and is thus constrained to the given structure and this way it may analyze it incorrectly, giving a poor performance to the predictions, as is for most of the predictions made (Table 7). As the method is structurally constrained and values are categorical, the method cannot consider subtle changes in the biomass or recruitment of the species, as they may fluctuate significantly, but not enough to be considered in another category, i.e. SSB fluctuating from high to medium high.

The method we used would most likely be at its best if a network structure was learned from a database or similar raw data bank with several, randomly selected species in the same geographical area with the TABU and HC algorithms without any whitelisting or prior bias given to the algorithm. With the learned networks, the nodes with no interactions would be dropped out, the models would be fitted and the model with the best score would be chosen. The strength of the arcs would be computed and lastly, the interactions that are the strongest would be used for recruitment and biomass predictions. This way the amount of data and researcher bias would be minimized, and the model would have a maximized predictive capability.

As a MSc thesis, this analysis could not be a most comprehensive one. Our results suggests, however, that with enough time and computing power, it would be possible to choose an area of interest, take the species in that area, often being dozens of species, and compute the interactions of those species based solely on the data available. There is potential to also use other sources of information, e.g. weather data, for computing other dependencies with the approach. This way there is a possibility to find interactions between species that have not been detected previously, be it environmental association or interactions due to predatory or foraging behavior. The high amount of ecosystem models available suggests that there is a belief among ecosystem modelers that useful causalities exist, outside of the RAM database approach applied here.

# APPENDIXES

| | TABU | | | | HC | | |
|---|---|---|---|---|---|---|---|
| loglik-g | from | to | strength | | from | to | strength |
| | ArcticHaddockR | ArcticPollockSSB | -8.944611 | | ArcticCodSSB | ArcticHaddockSSB | -48.416711 |
| | ArcticHaddockSSB | ArcticPollockSSB | -3.818640 | | ArcticCodR | ArcticHaddockR | -13.266583 |
| | ArcticCodSSB | ArcticPollockSSB | -2.163808 | | ArcticHaddockR | ArcticPollockSSB | -8.944611 |
| | ArcticHaddockSSB | ArcticCodSSB | -48.416711 | | ArcticHaddockSSB | ArcticPollockSSB | -3.818640 |
| | ArcticHaddockR | ArcticCodR | -13.266583 | | ArcticCapelinR | ArcticHaddockSSB | -2.364540 |
| | ArcticCapelinR | ArcticCodSSB | -3.340399 | | ArcticCodSSB | ArcticPollockSSB | -2.163808 |
| | | | | | | | |
| bic-g | from | to | strength | | from | to | strength |
| | ArcticHaddockR | ArcticPollockSSB | -7.1391523 | | ArcticCodSSB | ArcticHaddockSSB | -46.6112518 |
| | ArcticHaddockSSB | ArcticPollockSSB | -2.0131811 | | ArcticCodR | ArcticHaddockR | -11.4611237 |
| | ArcticCodSSB | ArcticPollockSSB | -0.3583493 | | ArcticHaddockR | ArcticPollockSSB | -7.1391523 |
| | ArcticHaddockSSB | ArcticCodSSB | -46.6112518 | | ArcticHaddockSSB | ArcticPollockSSB | -2.0131811 |
| | ArcticHaddockR | ArcticCodR | -11.4611237 | | ArcticCapelinR | ArcticHaddockSSB | -0.5590813 |
| | ArcticCapelinR | ArcticCodSSB | -1.5349399 | | ArcticCodSSB | ArcticPollockSSB | -0.3583493 |
| | | | | | | | |
| aic-g | from | to | strength | | from | to | strength |
| | ArcticHaddockR | ArcticPollockSSB | -7.944611 | | ArcticCodSSB | ArcticHaddockSSB | -47.416711 |
| | ArcticHaddockSSB | ArcticPollockSSB | -2.818640 | | ArcticCodR | ArcticHaddockR | -12.266583 |
| | ArcticCodSSB | ArcticPollockSSB | -1.163808 | | ArcticHaddockR | ArcticPollockSSB | -7.944611 |
| | ArcticHaddockSSB | ArcticCodSSB | -47.416711 | | ArcticHaddockSSB | ArcticPollockSSB | -2.818640 |
| | ArcticHaddockR | ArcticCodR | -12.266583 | | ArcticCapelinR | ArcticHaddockSSB | -1.364540 |
| | ArcticCapelinR | ArcticCodSSB | -2.340399 | | ArcticCodSSB | ArcticPollockSSB | -1.163808 |

**Appendix 1. Arc strength calculated for every arc for TABU and HC algorithms with loglik-g, bic-g and aic-g scores. The nodes in "from" are parent nodes and "to" are child nodes for the models.**

## ACKNOWLEDGEMENTS

Thank you to my supervisor, Sakari Kuikka, for an interesting and challenging topic, which made my hair gray, but also made me learn something that will be invaluable for me in the future.

Thank you to various fellow students that helped me with the R platform when I started the thesis, as I also started studying R at the same time.

And most importantly, thank you to my parents, Marja-Liisa and Tom, and to my godparents, Irene and Kari, for you continuous and irreplaceable support throughout my entire university career and especially these past two years.

# REFERENCES

Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., & Salmerón, A. (2011). Bayesian networks in environmental modelling. Environmental Modelling & Software, 26(12), 1376-1388.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Selected papers of hirotugu akaike (pp. 199-213). Springer, New York, NY.

Beretta, S., Castelli, M., Gonçalves, I., Henriques, R., & Ramazzotti, D. (2018). Learning the structure of Bayesian Networks: A quantitative assessment of the effect of different algorithmic schemes. Complexity, 2018.

Beygelzimer, A., Langford, J., Lifshits, Y., Sorkin, G., & Strehl, A. L. (2014). Conditional probability tree estimation analysis and algorithms. arXiv preprint arXiv:1408.2031.

Brooks, S. (1998). Markov chain Monte Carlo method and its application. Journal of the royal statistical society: series D (the Statistician), 47(1), 69-100.

Chen, Z. (2003). Bayesian filtering: From Kalman filters to particle filters, and beyond. Statistics, 182(1), 1-69.

Chickering, D. M. (2013). A transformational characterization of equivalent Bayesian network structures. arXiv preprint arXiv:1302.4938.

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. J. Mach. Learn. Res., 15(1), 3741-3782.

Corman, F., & Kecman, P. (2018). Stochastic prediction of train delays in real-time using Bayesian networks. Transportation Research Part C: Emerging Technologies, 95, 599-615.

Duodecim 2020: Summary – Bayesian reasoning in diagnostics of coronavirus infection, https://www.duodecimlehti.fi/duo15670#duo-comments-start (accessed 22.3.2021)

Heritage 2008: Economic Impacts of the Cod Moratorium, https://www.heritage.nf.ca/articles/economy/moratorium-impacts.php (accessed 20.04.2021)

European Commission Executive Agency for Small and Medium-sized Enterprises (EASME), 2018: Improving Cost-Efficiency of Fisheries Research Surveys and Fish Stocks Assessments using Next-Generation Genetic Sequencing Methods. Tender reference number EASME/EMFF/2018/015

Fenton, N., & Neil, M. (2011). The use of Bayes and causal modelling in decision making, uncertainty and risk. CEPIS Upgrade, 12(5), 10-21.

Froese, R., Demirel, N., Coro, G., Kleisner, K. M., & Winker, H. (2017). Estimating fisheries reference points from catch and resilience. Fish and Fisheries, 18(3), 506-526.

Fulton, E. A., Smith, A. D., Smith, D. C., & van Putten, I. E. (2011). Human behaviour: the key source of uncertainty in fisheries management. Fish and fisheries, 12(1), 2-17.

Gasparini, M. (1997). Markov chain Monte Carlo in practice.

Geiger, D., & Heckerman, D. (1994). Learning gaussian networks. In Uncertainty Proceedings 1994 (pp. 235-243). Morgan Kaufmann.

Gilbert, D. J. (1997): "Towards a new recruitment paradigm for fish stocks." Canadian Journal of Fisheries and Aquatic Sciences 54.4: 969-977.

Jewson, S., & Penzer, J. (2004). Optimal year ahead forecasting of temperature in the presence of a linear trend, and the pricing of weather derivatives. Available at SSRN 563943.

Koller, D., & Friedman, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.

Margaritis, D. (2003). Learning Bayesian network model structure from data. Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science.

Mitchell, T. M. (1996): Machine Learning. New York (NY): McGraw-Hill, 1997. Print.

Myers, R. A., Bridson, J., & Barrowman, N. J. (1995). Summary of worldwide spawner and recruitment data. Department of Fisheries and Oceans Canada Northwest Atlantic Fisheries Centre, St. John's, NF.

Myers, R. A. (2001). Stock and recruitment: generalizations about maximum reproductive rate, density dependence, and variability using meta-analytic approaches. ICES Journal of Marine Science, 58(5), 937-951.

Moran, B. M., & Anderson, E. C. (2019). Bayesian inference from the conditional genetic stock identification model. Canadian Journal of Fisheries and Aquatic Sciences, 76(4), 551-560.

Ogle, D. H. (2018). Introductory fisheries analyses with R. CRC Press.

Oniśko, A., & Druzdzel, M. J. (2014, June). Impact of bayesian network model structure on the accuracy of medical diagnostic systems. In International Conference on Artificial Intelligence and Soft Computing (pp. 167-178). Springer, Cham.

Patterson, K., Cook, R., Darby, C., Gavaris, S., Kell, L., Lewy, P., Stefánsson, G. (2001). Estimating uncertainty in fish stock assessment and forecasting. Fish and fisheries, 2(2), 125-157.

Portet, S. (2020). A primer on model selection using the Akaike Information Criterion. Infectious Disease Modelling, 5, 111-128.

Ricard, D., Minto, C., Jensen, O. P., & Baum, J. K. (2012). Examining the knowledge base and status of commercially exploited marine species with the RAM Legacy Stock Assessment Database. Fish and fisheries, 13(4), 380-398.

Russell SJ, Norvig P (2009). Artificial Intelligence: A Modern Approach. Prentice Hall, 3rd edition.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of statistics, 6(2), 461-464.

Kotsiantis, S., & Kanellopoulos, D. (2006). Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 47-58.

Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. science, 338(6106), 496-500.

THL 2021: Coronavirus tests, https://thl.fi/en/web/infectious-diseases-and-vaccinations/what-s-new/coronavirus-covid-19-latest-updates/symptoms-and-treatment-coronavirus/coronavirus-tests (accessed 22.03.2021)

Van de Schoot, R., & Miocević, M. (2020). Small sample size solutions: A guide for applied researchers and practitioners (p. 284). Taylor & Francis.

Walters, C., & Maguire, J. J. (1996). Lessons for stock assessment from the northern cod collapse. Reviews in fish biology and fisheries, 6(2), 125-137.

Worldometer 2021: Coronavirus in Finland, https://www.worldometers.info/coronavirus/country/finland/ (accessed 22.03.2021)

Yang, S., & Chang, K. C. (2002). Comparison of score metrics for Bayesian network learning. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 32(3), 419-428.

Yuan, C., & Druzdzel, M. J. (2006). Importance sampling algorithms for Bayesian networks: Principles and performance. Mathematical and Computer Modelling, 43(9-10), 1189-1207.

**R-package references:**

BNlearn: Journal of Statistical Software July 2010, Volume 35, Issue 3. http://www.jstatsoft.org/ Learning Bayesian Networks with the bnlearn R Package Marco Scutari University of Padova

ggplot2: H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

GGally: Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Jason Crowley (2021). GGally: Extension to 'ggplot2'. R package version 2.1.0. https://CRAN.R-project.org/package=GGally

bnviewer: Robson Fernandes (2020). bnviewer: Bayesian Networks Interactive Visualization and Explainable Artificial Intelligence. Rpackage version 0.1.6. https://CRAN.R-project.org/package=bnviewer

forecast: Hyndman RJ, Khandakar Y (2008). “Automatic time series forecasting: the forecast package for R.” _Journal of Statistical Software_, *26*(3), 1-22. <URL: https://www.jstatsoft.org/article/view/v027i03>.

**Copyrighted material:**

Map of major fishing area 27: http://www.fao.org/fishery/area/Area27/en (accessed 10.02.2021)

**ICES reports:**

ICES annual report 2019: https://www.ices.dk/sites/pub/Publication%20Reports/ICES%20Annual%20Report/Annual%20Report%202019.pdf (accessed 04.03.2021)

ICES cod 2016: Cod (Gadus morhua) in subareas 1 and 2 (Northeast Arctic), http://www.ices.dk/sites/pub/Publication%20Reports/Advice/2016/2016/cod-arct.pdf (accessed 12.1.2021)

ICES capelin 2016: Capelin (Mallotus villosus) in subareas 1 and 2 (Northeast Arctic), excluding Division 2.a west of 5°W (Barents Sea capelin), http://www.ices.dk/sites/pub/Publication%20Reports/Advice/2016/2016/cap-bars.pdf (accessed 10.02.2021)

ICES haddock 2016: Haddock (Melanogrammus aeglefinus) in subareas 1 and 2 (Northeast Arctic), http://www.ices.dk/sites/pub/Publication%20Reports/Advice/2016/2016/had-arct.pdf (accessed 10.02.2021)

ICES pollock 2016: Saithe (Pollachius virens) in subareas 1 and 2 (Northeast Arctic), http://www.ices.dk/sites/pub/Publication%20Reports/Advice/2016/2016/sai-arct.pdf (accessed 10.02.2021)